DOCUMENT RESUME

ED 274 192 FL 016 034

AUTHOR Mason, Victor W.
TITLE Item Banks: Benefits, Constraints and Security Systems.
PUB DATE Apr 86
NOTE 32p.; Paper presented at the Annual Meeting of the International Association of Teachers of English as a Foreign Language (Brighton, England, April 1-4, 1986).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Confidentiality; English (Second Language); *Item Banks; *Language Tests; Listening Skills; Multiple Choice Tests; Reading Skills; Second Language Instruction; Test Construction; *Test Format; *Test Items

ABSTRACT

Part I discusses the many benefits of having an item bank of four-option multiple-choice questions for test construction in English language instructional programs. The advantages of the item bank include: increased effectiveness, efficiency, and professionalism among teachers and administrators in test development, design, writing, review, and evaluation; greater control and rationale in grouping students according to their proficiency levels; and a more consistent program of high standards in test scoring, grading, and interpretation. Part II addresses the strengths and limitations of the four-option multiple-choice format, and related considerations such as item weighting, the length, difficulty, and timing of parts that affect test administration and item bank development. Part III outlines administrative procedures for use before and after test administration to preserve the security of the item bank. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Item banks: benefits, constraints and security systems

Victor W. Mason

ED274192

Abstract

Part I of this paper discusses the many advantages, to both administrators and teachers of an ELT program, of an item bank (IB) of 4-option multiple-choice questions intended to measure those linguistic variables for which objective assessment seems to produce highly valid results, such as the receptive skills and the components. Assuming sufficient release time from classroom instruction for testing administrators, several major program benefits from IBs should accrue: 1) markedly increased efficiency and quality in test-development efforts; 2) greater staff professionalism in the design, writing, review and evaluation of major program examinations; 3) more rational classroom enrollment patterns; and, thus, 4) significant growth in the administrative accountability of an ELT program.

Part II deals with the strengths and limitations of the 4MC format, to assist IB developers in knowing when use of this format is appropriate for measurement purposes. Also dealt with are related considerations like the weighting of items and the length, difficulty and timing of parts, all of which have an important bearing on test-administration results and thus on IB development. Part III is intended as an aid to IB developers in codifying the many administrative procedures staff members must follow, before and after test administrations, to prevent the security of the IB for any major program examination from ever being seriously compromised.

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Vic Mason

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL 016 034

Benefits

One of the most valuable components of an ELT program is the item bank of 4-option multiple-choice (MC) questions. When the MC format is used for those testing tasks for which it is appropriate, a great many benefits can accrue to a program from its development of an item bank (IB), primarily in two vital areas of concern: the assembly of new test forms and the maintenance of program academic standards.

Development of new test forms

The advantages of having available an item bank for test construction are several. It saves a teaching staff a tremendous amount of time in preparing new test forms (Popyuk 1980: 47) since MC item-writing is a very difficult and laborious process (Harris 1969: 8). Moreover, creation of an IB assumes the existence of a set of professional criteria specifying rational test-assembly procedures (Malcolm, 1980), allowing test developers to better control such basic variables as content, difficulty levels, length and reliability of each part and of the test as a whole.

On the basis of results of previous test administrations, the IB reveals the performance of each item, particularly its difficulty and discrimination indices, and specifies the relative effectiveness of each of the four options. When each question in the test is weighted equally at one point apiece, valuable part and total-test data then can also be readily obtained, including means, standard deviations, reliabilities and correlation coefficients.

Careful control of item and part characteristics in timed tests also makes possible more meaningful interpretation of students' test scores from semester to semester. This is because the "yardstick" employed in measuring students' knowledge, ability or achievement in English can be held relatively constant in major program examinations such as placement and achievement tests, even if parts and items are not strictly equated statistically. Means, standard deviations and reliabilities tend to remain quite consistent over time for similar student populations, thus enhancing the ability of administrators to predict group performance, roughly speaking, and to better maintain comparability of test effectiveness and program standards from one semester or course to another.

A major benefit resulting from the foregoing program activities can be growth in staff professionalism in testing matters (Woodford 1980: 97-8) and thus in program effectiveness. Staff skills are consolidated in crucial activities like test design, item-writing, test review and revision, and item and test analysis. At the same time, testing administrators improve their organizational and "clerical" skills, as they organize test data; develop IB cards; store raw data and the results of item and test analysis (e.g., answer sheets, computer printouts, etc.); prepare succinct and lucid memos to those responsible for collecting the test data; and finally prepare short summary reports in comprehensible language as feedback to the teaching staff.

A final benefit of obtaining and analyzing the empirical data on MC-test performance is that highly reliable objective parts assumed by the teaching staff to be valid measures of student proficiency, knowledge or progress then become standards for determining the effectiveness of subjectively scored tasks included in the same test form for the stated purpose of assessing the same linguistic variables. For instance, teachers commonly include productive or mixed-mode tasks to test students' abilities in the components and the two receptive macro-skill areas. They often justify use of such tasks and formats because of their honest belief that since they <u>think</u> a test part measures what they

claim it does, it thereby has face/content validity (Stevenson 1981: 269). In design it might even appear to be excellent. But if students' scores on such tasks, especially when done by homogeneous NL groups, fail to correlate sufficiently 1) with scores on well-written and reliable same-trait objective test parts (as when scorers are inadequately trained to apply similar scoring criteria) or 2) with key examination and course criteria (such as test total scores, course total marks and course final grades), then the subjectively scored parts may very well produce invalid part scores, and course grades might turn out to be unfair. Generally speaking, however, the results of reliably scored subjective parts can be expected to correlate satisfactorily with those of well-prepared objective parts (Harris 1969: 70).

Maintenance of program standards

The IB of MC items is thus clearly of central importance to every ELT program which desires to assure consistently high standards in test development and in scoring and grading practices. In addition, however, it can have a considerably broader beneficial impact at the program level---by promoting administrative accountability. Without test data obtained in a scientific manner, administrators cannot know 1) if major program tests are consistently effective, 2) if classroom instruction is generally effective for all staff members, and 3) if students are making any significant progress in English.

Analysis of records of MC item, part and test performance can provide the empirical foundation upon which a program obtains the baseline data required for rudimentary in-house research and evaluation (Tucker 1980) in these three key areas of concern to educators in all types of programs. Teachers along with administrators can thus learn in a limited but very meaningful and practical way to help collect and analyze raw test data and to see themselves as experimenters (Shaw, et. al., 1983: 132)---a role which, to their frequent surprise, they take to with enthusiasm when not intimidated by tables, graphs, formulas and obscure statistical terminology (Alderson and Hughes 1981: 5-7).

One of the first and most valuable spin-offs of the baseline data records obtained from the above-mentioned test-development activities should be the increasingly homogeneous grouping of new students by ability level through the use of highly reliable 100-per cent objective placement tests (Mason 1984). A principal frustration of EL teachers and students almost everywhere is the enrollment of students of widely differing English abilities into the same course levels and classrooms (Early and Bolitho 1981: 118-22). Responsibility for suitable enrollment patterns is primarily, if not entirely, administrative in nature; with the powerful tool of the IB of 4MC questions plus the record of test analysis based upon it, administrators can obtain the

vital baseline data they need in order to do their best to help rationalize classroom enrollment patterns---and thus to markedly help teachers teach and students learn (Ostojic 1981).

Control of placement policy and practice, affecting new---i.e., unknown---students, is only one half of the enrollment problem, however. The more troublesome "half" is control of policy and practice in connection with the administration of major course achievement tests like midterm and final examinations and the assignment of course grades to continuing---i.e., known---students. Teachers of the same program often have surprisingly different standards in the assignment of test and course marks and grades. The only way this problem can be overcome is for a program's administration to establish guidelines for the assembly of parts of major tests; to see that subjective test parts are panel-marked according to well-defined scoring criteria; to see that test and course grades are assigned with reference to the scores of students of all sections of the same course; and to be sure that the nebulus area of "coursework" neither constitutes too large a part of the course mark (a maximum of 20-25 per cent seems fair to the diligent and strict teacher) nor is seen by overly lenient staff members as a means of "padding" students' semester marks with undeserved points.

When an ELT program administration sees to it that the parts of major achievement tests validly employing the MC format are assembled according to the same strict procedural principles as are placement tests, that program has taken a giant step forward in controlling the factors most probably having the greatest influence on the degree to which its sequence of courses demonstrates rational patterns of enrollment (see Hudson & Lynch 1984:176).

In particular, it is important that difficulty levels of placement tests and final examinations be comparable. MC test parts written at "middle difficulty" (Cooper 1972: 341) for a target group are generally quite effective in spreading students' scores out, and tend on the same test form to yield lower scores than do subjectively scored test parts. For example, it is the unusual examination on which mean scores on MC reading or listening parts are higher than those for speaking or writing, assuming that defensible criteria of test design and scoring have been followed. IBs for the objective parts of placement tests and final examinations greatly help test developers to assemble such different major program tests according to the same standards of difficulty and discrimination. Since questions from the different IBs[1] must be based upon the same program curriculum (i.e., students taking placement tests and final

examinations will be enrolled subsequently into the same courses), the IBs of the MC questions constitute the indispensable practical basis for rationalizing placement and promotion decisions, as total scores obtained from the two types of examinations can be given a common theoretical interpretation, and enrollment decisions based upon them, a ready justification.

Finally, for those ELT programs which desire to have a clearer idea of student progress in English, a series of administrations of valid standardized tests like the TOEFL can help to define that growth on both an individual and a group basis (Swinton 1983). A program's effectiveness can thus be increasingly monitored by its administrators, as they concentrate ever more on the real challenges of "professional" (or qualitative) administration as opposed to the many "bureaucratic" (or quantitative) aspects of running a program, such as assigning rooms, setting schedules, ordering books, meeting payrolls, supervising student and teacher attendance, and shuffling papers, which an executive assistant with organizational talents could be hired to perform.

Related considerations

If a program is to truly benefit from the development of IBs, certain prerequisite conditions need to be met. The most obvious immediate consequence of all of the above-mentioned test-development and program-evaluation activities is a significant increase in workload on all of those staff

members assigned test-related duties. Testing administrators burdened with such responsibilities must be given considerable release time from teaching duties to permit the necessary tasks to be performed.

The testing administrator conscientiously performing all incidental tasks required is extremely busy. His responsibilities include: obtaining comprehensive and coherent statements of program and course curricular objectives for program files; designing, or giving guidance in the design of, test content specifications based upon the stated course objectives; giving test-writing assignments; coordinating test-development activities such as chairing meetings to resolve staff differences of opinion on pre-test results; supervising the typing and doing proofreading of each typed draft of the test (preliminary and final) during the development or assembly process; and taking responsibility for the final form of the test prior to reproduction. Thereafter, he becomes involved in much dull but essential "busywork" including overseeing the printing, collating and stapling of test pages, and the checking, numbering and packaging of test booklets; is responsible for test security at all stages of the development process; explains and supervises the carrying out of program procedures ("rules") of test administration; arranges for the secure storage and expeditious destruction of used test booklets; arranges for the analysis of test results; oversees the updating of the IBs and the

testing files; prepares feedback to staff members in the form of short reports; and prepares the necessary inter-staff communications throughout the term or semester that are required to keep administrators and teachers on schedule in, and aware of, the performance of all of the foregoing test-development matters. In short, full release time from teaching, in such circumstances, may be advisable.

The second precondition required for a program to reap the full benefits of its IBs is to have available the services of certain specialized institutional staff and facilities. For the analysis of the results of tests, it is best to have staff members with some understanding of basic statistical procedures to perform item and test analysis and, if necessary, to provide liaison with computer programmers.

It can be very laborious to obtain basic item, part and test statistics "by hand." However, for the many EL teachers with the desire but lacking the technical background and/or program support and facilities, IB-development can nonetheless proceed according to the guidelines of such useful references as Diederich (1973:2) for item analysis by the "high-low" method; Harris (1969: 142-5), for the calculation of correlations and the estimation of reliabilities: and Downie and Heath (1974:106), for the abac to obtain approximate item-discrimination indices.

Finally, certain reservations need to be mentioned in the development of IBs. First, it is not appropriate to use the MC format for all EL testing tasks; for many testing purposes, especially in truly communicative situations, it lacks face/content validity. Second, there must be strict observation of IB security systems; otherwise, much valuable staff work can be quickly compromised. Parts II and III of this paper, respectively, deal in detail with these two important considerations.

Strengths and limitations of the multiple-choice test

Questions about the format

Since this paper advocates the use of item banks of 4-option MC questions, some attention is given to the legitimate doubts often expressed about this item-type. The area of proficiency testing tends to be one of the more controversial topics in the field of ELT (Farhady 1979: 347), 347), and it is hard to think of any that is more so than the subject of the MC format. Nevertheless, the 4-option MC item is used almost universally in EL programs, and a survey might well find it the single most commonly employed type of question, whether in ELT or in other categories of academic programs.

Harris (1969:8) says the fact that "the examinee does not have to think of his own answers" is the single most

frequent criticism of the format. Though a student might not be able to recall the answer on his own, he may nevertheless recognize it when it is presented as an option and so receive credit not possible in a supply-type test (Valette 1967: 36).

Second, it is generally agreed that the MC item is not valid for the measurement of the productive macro-skill areas of speaking and writing (Clark 1972: 40). For these skills, direct-proficiency performance tests that have students actually speak and write are employed to provide genuine face/content validity (Clark 1978: 24).

A third common criticism of the MC format focuses in particular on discrete-point items, such as those of grammatical structure and vocabulary, which are said to provide too little context to constitute samples of authentic language use. Thus, B. Carroll (1980:9) cautions that "detaching test items from their communicative context is to risk finding little about the learner's behavior which is not trivial," and adds (Ibid.: 35): "The MC test has few of the characteristics of normal communication... [and so may lack] relevance and authenticity."

Related to the foregoing is the observation that MC tests tend to assess language variables in isolation from one another (Clark 1972: 32-3), when typical language use involves the simultaneous interaction of many, even

most, of the components and skill-areas. As a result, J. Carroll as long ago as 1961 recommended that EL proficiency tests be designed to include more "integrative" tasks that place greater emphasis on "the total communicative effect of an utterance" (J. Carroll 1961: 318).

A final problem with MC tests is the fact that they "are extremely difficult to prepare" (Oller 1976:22). As a consequence, "very considerable skill and time are required" for their preparation (Harris 1969: 8)---a drawback seeming, however, to reflect less on the items than on those responsible for producing them.

Advantages of the multiple-choice format

On the other hand, the MC format has many important practical advantages, including "objectivity of scoring and economy in marking," which can "make it a useful component of a test system...(for) a narrow range of purposes" (B. Carroll 1980: 35). The extent of this usefulness depends on the kinds of test tasks concerned. When the use of the MC format is appropriate, it "is usually regarded as the most valuable and most generally applicable of all test forms" (Stanley 1965: 222). And, "other things being equal, MC items tend to provide the highest quality items" (Gronlund 1968: 26).

While the MC format may be widely regarded as invalid for the direct testing of the productive skills, it is considered very appropriate for well-designed and -written

integrative tasks assessing ability in the two receptive skill-areas of listening and reading comprehension. Such tests are usually seen as having high face/content validity, although the form of student response may initially appear a bit artificial. The task of putting an essentially non-linguistic mark on a piece of paper, however, assures that in testing listening and reading ability, scorers will not then be influenced by the effect of extraneous variables like the ability to write or the control of grammatical structure (Valette 1967: 173). Moreover, the test results are broadly suggestive of students' proficiency in the corresponding productive skills of speaking and writing, for as Woodford (1980:99-100) notes: "We know that performance in the receptive skills often correlates highly with performance in the productive skills", (Also, see Pike (1979:81-2) on the rationale behind the test content specifications of the TOEFL.)

The limited context of short MC items testing grammar and vocabulary can perhaps be faulted, but it is also a strength, in that a great many specific, important structural and lexical points may be sampled within a relatively short time. Since the MC format is especially useful in examining knowledge (Valette 1967: 173), the examiner can interpret the results of the test parts, each composed of many of these "mini-tests," with great accuracy. Either the subject knows the grammar and vocabulary or he does not. Moreover, the imaginative item-writer will avoid writing sentence-level items that

are meaningless or nonsensical. It is true that many sentence-level items used in actual classroom lessons and in tests lack sufficient "indexical" meaning (Widdowson 1983:101)---i.e., instantly recognizable reference to the real world---but with a bit more care, teachers could readily prepare much improved sentence-level teaching and testing materials.

Thus, separate MC parts testing the components and the receptive macro-skill areas should correlate reasonably well, since they may well be testing the same basic underlying abilities (Farhady 1979; 354). However, Farhady (Ibid., 354-6) cautions against inclusion of only discrete-point or only integrative measures in major examinations whose results are to be used for such important administrative purposes as making placement decisions. This is because individual examinees may perform quite differently on these two types of tasks, "regardless of their overall proficiency but depending on their previous educational and linguistic backgrounds" (Ibid., 355). For example, a study of the influence of native-language background upon results on the various discrete-point and integrative MC tasks of the TOEFL "revealed significant differences on seven-eighths of the TOEFL items" (Alderman and Holland 1981: iii).

Ducroquet lists several very important benefits of the MC item-type for the classroom teacher: The format permits a great many teaching points in a course syllabus

to be tested within a very short time. Major course achievement tests, therefore, may contain questions sampling the full range of course content and difficulty levels for each linguistic variable to be assessed. The great majority of students (i.e., those of "average" ability or better for a given course level) are expected to be able to respond to all items in a suitably timed test; whereas in a supply-type exam, they might be able to avoid revealing what they do not know or have not learned. With a large number of test questions, reliability tends to be high, and scores usually discriminate well among students of markedly different abilities (Ducroquet 1980: 242-3). Furthermore, as noted above, if subjective test parts are also reliably scored, correlations between the objective and the other test parts are generally quite satisfactory (Harris 1969: 70), resulting in fair scoring, grading and promotion decisions with which students will usually concur.

The criticism that the MC format artificially separates the testing of the linguistic variables (Clark 1972: 33) can also be looked at in another way. As J. Carroll (1961: 319) points out: "In some ways, a good test is like an experiment," in that the examiner should attempt to measure basically one skill or area of knowledge at a time. He adds, however, that "there is no reason why a variety of techniques should not be used in one examination"---including discrete-point and integrative MC parts,

along with supply-type parts and even multi-mode tasks when it is "necessary to test two or more things simultaneously" (Ibid., 320; also, Harris 1969: 71).

Some important considerations in MC test development

To produce good MC items for program IBs, however, it is very important that test writers follow the guidelines of good reference works on item-writing procedures. Errors common in test-writing, with special implications for the MC format, are discussed at length in Dieterich, et. al. (1979), and Henning (1982).

test difficulty

The common complaint that subjects need not think up their own answers in taking MC tests is a relevant one but not important if the item-writer always uses at least four options and makes all four of them equally attractive or plausible to a student who does not know the answer. The mean chance score on such MC parts is then only 25 per cent, so there is hardly a reward on a major test for guessing, unlike with 3-option or true-false questions, on which the chance score rises to 33 and 50 per cent, respectively. (See Stansfield 1985: 143.)

Once the IB grows for each relevant variable, average part difficulty from test to test is better controlled and made more consistent. A suitable target level for a particular group of students is what is called "middle difficulty"; that is, a test is prepared so that the

group average score will be around 62.5 per cent, because the subjects know about 50 per cent of the answers and guess approximately one fourth of the remainder. Test results should then help to produce a fairly wide range of total scores, allowing better discrimination between students of clearly different language abilities, to assist in such important administrative tasks as making admissions, placement, grading and promotion decisions. Furthermore, good MC parts of appropriate length will often work much more effectively than subjectively scored test parts both in being on average more challenging to most students, as determined by part-test results; and in spreading their scores out for more meaningful discriminations as to relative proficiency or achievement.

weighting of items

Weighting of MC items also has an important bearing on IB development. Calculations of test statistics and comparisons of relative item effectiveness are greatly facilitated by giving all MC items in a test equal weight, namely, one point each. Teachers commonly assign a weight to the items of some parts, such as reading and listening, some two, three and even four times greater than that of items in other parts, especially questions in the discrete-point tasks, on the ground that the former parts are intrinsically more difficult and should thus count more---in the absence of any statistical support for the claim.

In fact, many grammar and vocabulary questions turn out to be more difficult than those based on comprehension passages. Existence of the IB provides a priori evidence of an item's history of usage with comparable groups of subjects. Moreover, when any items that have arbitrarily been given a disproportionate weight turn out to be badly written or for any reason ineffective, their principal impact on test results is to disproportionately lower the reliability of that part of the examination (Malcolm, personal communication). Errors by teachers in test design or scoring then may well turn out inequitably to affect placement, grading or promotion decisions based on test results or course marks that happen to fall very near crucial cut-off points.

length of MC test parts

A final very important consideration in MC test design with a direct bearing on IB development is the determination of the appropriate length of a given MC part. Along with relative difficulty, item discrimination power is a vital item statistic and is directly determined by comparing the performance of each individual question with that of the entire test part in which it is found. The more items that are included in a test part, and the larger the number of subjects tested, the greater the data base on which item and part discrimination power are calculated.

Length and difficulty levels of major course achievement tests are frequently the subject of heated debate in ELT programs at test-development time. Teachers often differ widely in their views on the subject, such as during meetings to design and produce midterm and final examinations. Those with strict standards tend to insist on long, comprehensive and challenging test parts and make few friends among those who do not. (The problem seldom arises in the development of new placement tests, affecting as they do only unknown students.)

Unfortunately, published test-writing manuals say very little on the subject of MC part length to guide program administrators and classroom teachers. However, it is probable that MC test parts with fewer than 20 questions for a particular linguistic variable will produce an inadequate data base for purposes of IB development; the same may be true if the number of subjects falls much below around 20 individuals (Malcolm, personal communications). Conversely, the more items a well-prepared test contains, the higher the reliability it can be expected to achieve.

For end-of-course assessment purposes, it seems not unreasonable that test parts for structure and vocabulary be timed---either overtly or implicitly---at an average rate of not less than 2-3 items per minute. Especially for major course achievement tests, the student who has studied and prepared properly will tend to know the answers and proceed through the parts

rather quickly. Thus, diligent students of average ability at a given course level could arguably be presented, say, some 50-60 items of structure and the same number of vocabulary items and be given a total of 40 minutes for the completion of the two parts. On a 2½-hour (150-minute) final examination, three-fourths of test time would still remain for the paper-and-pencil assessment of other essential linguistic variables.

Listening and reading comprehension parts are more difficult to prepare and time. Quality of test results and IB development depends first and foremost upon the selection of comprehension passages of appropriate length, content and difficulty. Assuming that suitable passages have been obtained, as based on selection criteria found in good reference manuals, timing considerations will be quite different in the two receptive skills.

A "self-timed" listening comprehension test part of some 30 items in which the tape is never stopped, and the material is heard only once, could probably be completed within 20 minutes. This assumes that students do not have to spend an unreasonable amount of time absorbing stimulus material or therefore to depend excessively on memory to answer the questions.

Reading can be tested with a series of short passages of varying content and should involve a high ratio of questions asked to words read. With appropriately chosen

topics, a good reading test can yield at least one good comprehension question for every 20-25 words of text. It is then not unfair to set a time in minutes for that part of the test equivalent to the number of questions asked; e.g., 30 minutes for 30 items. However, given the predictable resistance of many teachers to such a strict standard, especially on final examinations, it might be advisable for testing administrators to be willing to allow students up to 50 per cent more minutes than there are questions on the reading part of an examination.

Either listening or reading can be tested quickly and efficiently with meaningful stimulus material ranging in length from one short sentence up to around 100-120 words, for listening, or up to 120-150 words, for reading. Short and long conversations are suitable to test both receptive skill-areas, as are short and long talks for listening and standard comprehension passages of various writing styles, topics and length for reading. Above all, it is vital that a teaching staff be urged to follow closely the guidelines on test development for comprehension parts of examinations, as found in good testing references. Much progress in the development of teacher-prepared tests in these two skill-areas would result if EL teachers avoided equating the notions of "communicativeness," "sufficient context" and "appropriate difficulty level" with the presentation of many, many hundreds of words of text, most of them irrelevant to the testing tasks specified, followed by a few MC or other questions

of excessive generality and specificity. Unfortunately, just such a flawed presentation is what teachers find as a model in all too many of their texts teaching and testing student comprehension in these two skill areas.

Assuring item-bank security

The difficulty and sheer laboriousness of preparing MC test parts of suitable length, difficulty and quality points up one of the IB's chief virtues: once in place, it saves a hard-working instructional staff a tremendous amount of time in developing new tests. With a sufficient number of items available for testing each linguistic variable, former students will be unable to help current ones to anticipate which specific questions are likely to be asked on particular learning points and which comprehension passages are likely to be included in a test.

In contrast with its labor-saving advantages, the IB of MC items has an unfortunate drawback. Since items are to be re-used in the future, loss of a single test booklet can compromise, and perhaps wipe out, a very sizable part of a program or course item bank. Thus, attention to security procedures is necessary when an IB is established.

Basic Considerations

In creating item banks, a teaching staff must appreciate that a lot of extra time and effort are required to assure the security of an IB. However,

scrupulous observance of the necessary procedures involved will repay their efforts many times over through the enormous reduction of test-development time required (especially for placement tests and course final examinations) each semester of each academic year, as well as through the significant increase in test quality and effectiveness that should soon follow.

Probably the single most useful procedure for a program staff to follow is quite simply never to discuss test-security matters with students. For example, a test booklet will have far less value to students (both literally and figuratively speaking) if they are kept unaware of the existence of such an entity as an "item bank." Thus, the program's first line of defense of its IBs is a judicious combination of student ignorance and staff reticence on the subject.

Furthermore, a program will need a separate IB for each type of major examination. Thus, for example, one IB will be needed for the placement tests given at the start of each term or semester. A separate IB will also be needed for each course final examination.[2] The major tests in a program all have unique roles, with their varying curricular emphases and content specifications.

Moreover, by not mixing items prepared for different tests, no more than one IB would be affected in the event of a significant breach of test security.

Useful procedures to follow

To keep an item bank secure, a number of important rules need to be observed in the test-development cycle of production, storage, administration and destruction of test booklets. First, waste and extra loose pages resulting from the printing, collating and stapling stages of production must be collected and destroyed under secure conditions. Second, the test production rooms should be placed off limits on a permanent basis to all students and unauthorized staff, as with any other confidential work place.

Next, both a test booklet and its answer sheet(s) must be numbered. Fourth, booklets should be packaged as soon as possible in large envelopes with the range of numbers of enclosed test booklets written on the outside. The packaged tests are then securely stored until the date of administration.

Sixth, in the classroom students should be permitted to take nothing to their seats except writing implements. In this way, they are unlikely 1) to absentmindedly pick up the test booklet with their own belongings at the end of the testing time and 2) to go unobserved by room proctors if they attempt to walk away with the booklet.

Seventh, the number of students in the room should be counted while they are taking the test. Students who finish and leave early one at a time present no problem, but those who remain until the time is up should be kept in the room until all test answer sheets have been collected and counted. If possible, test booklets should be arranged in numerical order and double-checked. The booklets should then, as a further precaution, be checked once more afterwards under less hectic circumstances.

Following the test administration, the booklets must be repackaged and stored under the same stringent conditions as before the test is given. Test destruction should be scheduled as rapidly as is practicable. Alternative methods include burning, burying and shredding. Not recommended is selling the test pages to friendly vendors at the local night market who wrap sweets and other delightful morsels in them for their legions of appreciative customers. (This writer has on occasion chanced to become reacquainted with some of his items in just this way, months after test administration.)

Further, the item cards in the IB should be stored in a locked filing cabinet, with all items of a particular linguistic category kept together, as in a box or with rubber bands. Similarly, the test master pages and file copies of the test booklets must be stored under similarly secure conditions. (Advancements in word-

processing and video-disc technology promise to make such record-keeping and filing tasks much more efficient for those programs that are able to utilize its benefits.)

Finally, students may wish to study and then debate their test results. This is far less likely in the case of tests in which no grades are involved, such as placement tests. It is highly probable, however, that many students will want to know the results and to see the questions of major course achievement tests. Unless a course has an item bank (with or without item and/or part statistics on file), there is no good reason for not simply returning the test booklets and student papers to the class. (This helps to avoid the widespread and undesirable practice of teachers re-using old tests over and over again simply because it is a lot of work to prepare new ones.)

If, however, an IB is developed for a test, obviously the test booklets can not be returned to the students. Considering the special importance everywhere attached to final examinations because of their weight in calculating total semester marks and thus in determining final grades, it is best if item banks are developed for them and that students be permitted, if they so wish, to make appointments to examine their test papers and results one at a time in the teacher's presence, without making notes.

As for major pre-final course common examinations (like midterms), it is probably desirable not to even attempt to create IBs. One reason is that, in the middle of a semester, a teacher is likely to want to review the test and its results in class with all students simultaneously. Moreover, end-of-class commotion makes it much more likely that in some classrooms of the course, test booklets might be carried off and disappear before the loss was discovered. Thirdly, however, a more positive reason for not developing IBs for midterm exams is that teachers need regularly to hone their test-development skills, both 1) "vertically"---i.e., through all the essential steps of reclarification of course objectives, design of test specifications, writing, pretesting, review, evaluation of results---and 2) "horizontally": i.e., through the practice of preparing test parts and items for each of the various language components and skills, whether tested separately or together and whether in objectively or subjectively scored test tasks. Fairly early in the semester, when teachers' energies are still far from being sapped, seems a good time to keep these vital professional skills alive and well.

---

[1]Each major program test requires its own IB.

[2]Item banks are not here recommended for midterm examinations or other major pre-final course progress or achievement tests, for reasons discussed in the last paragraph of the paper.

References

Alderman, D.L. and P.W. Holland. 1981. Item performance across native language groups on the Test of English as a Foreign Language (TOEFL research report No. 9). Princeton, N.J.: Educational Testing Service.

Alderson, J.C. and A. Hughes (eds.). 1981. Issues in language testing: ELT documents No. 111. London: The British Council.

Brown, H. Douglas. 1980. Principles of language learning and teaching. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Carroll, B.J. 1980. Testing communicative performance. Oxford: Pergamon Press.

Carroll, J.D. 1961. 'Fundamental considerations in testing for English language proficiency of foreign students' in Allen and Campbell (eds.) 1972.

Clark, J.L.D. 1972. Foreign language testing: theory and practice. Philadelphia: The Center for Curriculum Development, Inc.

-----. 1978. 'Psychometric considerations in language testing' in Spolsky (ed.) 1978.

Cooper, R.L. 1971. 'Testing' in Allen and Campbell (eds.) 1972.

Diederich, P.B. 1973. 'Short-cut statistics for teacher-made tests.' Princeton, N.J.: Educational Testing Service.

Dieterich, T.G., et.al. 1979. 'A linguistic analysis of some English proficiency tests.' TESOL Quarterly 13/4: 535-550.

Downie, N.M. and R.W. Heath. 1974. Basic statistical methods. New York: Harper & Row, Publishers.

Ducroquet, L. 1980. 'Objective or subjective testing? System 8/3: 237-243.

Early, P. and R. Bolitho. 1981. 'Reasons to be cheerful, or helping teachers to get problems into perspective...'. System 9/2: 113-124.

Farhady, H. 1979. 'The disjunctive fallacy between discrete-point and integrative tests.' TESOL Quarterly 13/3: 347-357.

Gronlund, N.E. 1968. Constructing achievement tests. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Harris, D.P. 1969. Testing English as a second language. New York: McGraw-Hill Book Co.

Malcolm, D.J. 1980. Test development manual. Kuwait University Language Center.

Mason, V.W. 1981. 'Creating an item bank.' Al-Manakh 5/1: 73-86.

-----. 1984. 'Using multiple-choice tests to promote homogeneity of class ability levels in large EGP and ESP programs.' System 12/3: 263-271.

Oller, J.W. Jr. 1976. 'Language testing today.' English Teaching Forum.

Ostojic, B. 1981. 'Streaming in foreign-language study.' English Teaching Forum 19/2.

Pike, L.W. 1979. An evaluation of alternative item formats for testing English as a foreign language (TOEFL research report No. 2). Princeton, N.J.: Educational Testing Service.

Popyuk, W. 1980. 'A model for an item bank in second language proficiency testing.' System 8/1.

Shaw, P., et.al. 1983. 'Talking shop: the L.A. tapes.' ELT Journal 37/2: 129-137.

Stanley, J.C. 1965. Measurement in today's schools. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Stevenson, D.K. 1981. 'All of the above': on problems in the testing of FL reading.' System 9/3: 267-273.

Swinton, S.S. 1983. A manual for assessing language growth in instructional settings (TOEFL research report No. 14). Princeton, N.J.: Educational Testing Service.

Tucker, G.R. 1980. Comments in speech delivered to KULC staff and Kuwait educational officials at the Adeliya Campus of Kuwait University, May 1980.

Valette, R.M. 1967. Modern language testing: a handbook. New York: Harcourt Brace Jovanovich, Inc.

Widdowson, H.G. 1983. 'Competence and capacity in language learning.' On TESOL '82: 97-106.

Woodford, P.E. 1980. 'Foreign language testing.' The Modern Language Journal 64/1: 97-100.